فصلنامه علمی - پژوهشی فرماندهی و کنترل، سال دهم، شماره یک، بهار ۱۴۰۵

ارائه روشی برای افزایش تمرکز شبکه عصبی عمیق بر سوژه اصلی مبتنی بر توضیح‌پذیری برای مدل‌های طبقه‌بندی

علی جعفری[1]، شریف حسنی[2]*

تاریخ دریافت: ۱۴۰۵/۰۳/۱۱ **تاریخ پذیرش: ۱۴۰۵/۰۶/۲۳**

چکیده

با پیشرفت استفاده از هوش‌مصنوعی و یادگیری‌عمیق در سال‌های اخیر در صنایع حساس و حیاتی مانند سامانه‌های فرماندهی و کنترل نظامی، وجود عناصری مانند قابلیت اعتماد، توانایی تفسیر نتایج و درک ساختار تصمیم‌گیری مدل‌های آموزش داده شده بیش از پیش احساس شده است. به همین منظور راهکارهایی مانند الگوریتم‌های توضیح‌پذیری مدل مانند Grad-CAM، SHAP، LIME و... ارائه شده‌اند. پس از آن در سال‌های اخیر رویکردی برای افزودن راهنمایی‌های توضیح‌پذیری در طی فرآیند آموزش مدل با نام اختصاری EGL نیز مطرح شده است. که هدف آن افزایش دقت پیش‌بینی و کسب اعتماد کاربر نسبت به نتایج مدل است. در این پژوهش راهکاری نوین با استفاده از رویکرد EGL برای افزایش تمرکز مدل‌های طبقه‌بندی بر سوژه اصلی ارائه شده است. تابع هدف پیشنهادی دارای جملاتی برای طبقه‌بندی، تمرکز بر سوژه اصلی و کاهش توجه به نقاط فاقد اهمیت در تصویر است. روش پیشنهادی بر روی شبکه ResNet۱۸ با مجموعه داده TinyImageNet آموزش مجدد داده شده است و نتایج آن از دو جنبه دقت پیش‌بینی و کیفیت توضیح‌پذیری ارتقا یافته است. همچنین در مقایسه با پژوهش‌های پیشین روش پیشنهادی موجب افزایش تمرکز شبکه عصبی بر روی سوژه اصلی و کاهش تمرکز مدل بر ویژگی‌های پس‌زمینه شده است.

واژگان کلیدی: شبکه‌های عصبی عمیق، توضیح‌پذیری، تفسیر پذیری، یادگیری مبتنی بر توضیح‌پذیری **EGL**، طبقه‌بندی تصاویر.

[1] نویسنده مسئول، استادیار گروه علمی هوش مصنوعی، مجتمع دانشگاهی برق و رایانه، دانشگاه صنعتی مالک اشتر iustuser@mut.ac.ir
[2] پژوهشگر دانشگده مهندسی برق و کامپیوتر، دانشگاه صنعتی مالک اشتر

[Downloaded from ic4i-journal.ir on 2026-11-08]

۱. کلیات

هوش مصنوعی به عنوان یکی از پیشرفته ترین و تاثیرگذارترین فناوری های نوین با ورود به عرصه های نظامی به ویژه سامانه‌های فرماندهی و کنترل، تحولی بنیادین را رقم زده است. سامانه‌های فرماندهی و کنترل مدرن با بهره گیری از توان تحلیلی و تصمیم گیری سریع هوش مصنوعی وارد مرحله ای جدید از خودکارسازی و دقت بالا شده اند. ولی با توجه به حساسیت بسیار زیاد سامانه‌های فرماندهی و کنترل نظامی، عدم شفافیت موجود در شبکه‌های عمیق مورد استفاده در هوش مصنوعی، باعث غیر قابل اعتماد شدن این فناوری در این حوزه شده است. در این راستا قابلیت توضیح‌پذیری در مدل‌های یادگیری ماشین به یکی از محورهای اصلی پژوهش‌های هوش‌مصنوعی مدرن تبدیل شده است.

معمولاً روش‌های توضیح‌پذیری به دو دسته‌ی روش‌های پس‌توضیحی[۱] و ذاتا توضیح‌پذیر[۲] تقسیم می‌شوند [۱]. در روش‌های پس‌توضیحی، تصمیمات مدل از طریق تحلیل اهمیت ویژگی‌ها پس از اتمام فرآیند آموزش تفسیر می‌شوند [۲]. با این حال، این روش‌ها عمدتاً فقط به تحلیل پیش‌بینی‌های مدل می‌پردازند و تأثیری در بهبود عملکرد یا فرآیند یادگیری مدل ندارند. در پاسخ به این محدودیت، رویکردی به نام یادگیری مبتنی بر توضیح‌پذیری[۳] (EGL) معرفی شده است که در آن توضیحات ارائه شده توسط روش‌های پس‌توضیحی مستقیماً وارد فرآیند یادگیری می‌شوند. این کار منجر به ساخت مدل‌هایی می‌شود که ذاتاً قابل تحلیل هستند و بر ویژگی‌های معنایی و قابل‌اعتماد تمرکز دارند. این روش‌ها با هم‌تراز کردن هدف آموزشی با استدلال‌های قابل‌درک برای انسان، به‌طور هم‌زمان قابلیت تحلیل و عملکرد مدل را در وظایف پایین‌دستی بهبود می‌دهند [۱].

با این حال، یکی از چالش‌های اساسی در یادگیری مبتنی بر توضیح‌پذیری EGL، نبود یک چارچوب ساخت‌یافته برای هدایت شبکه عصبی به سمت تمرکز هدفمند بر سوژه‌ی اصلی تصویر و استخراج ویژگی‌های کلیدی، متمایز و معنادار است. این چالش به‌ویژه در مجموعه‌داده‌هایی که کلاس‌ها دارای هم‌پوشانی بالا در فضای ویژگی هستند، نمود بیشتری پیدا می‌کند. در چنین شرایطی، کنترل فرآیند یادگیری مدل بر اساس اهداف توضیح‌پذیری به‌درستی انجام نمی‌شود و شبکه تمایل دارد به سرنخ‌های زمینه‌ای یا نویزهای مشترک میان کلاس‌ها اتکا کند. به‌عنوان نمونه، در مجموعه‌داده‌ای شامل کلاس‌های «قورباغه دم‌دار» و «قورباغه گاوی»، ویژگی‌های استخراج‌شده توسط لایه‌های کانولوشنی به‌شدت به یکدیگر نزدیک بوده و تمایز مؤثر میان این دو کلاس برای شبکه‌های عصبی به فرآیندی پیچیده تبدیل می‌شود.

رفع این مسئله مستلزم تقویت مؤلفه‌ی هوش[۴] از طریق بهره‌گیری از فراداده‌های حقیقت مطلق[۵] توضیح‌پذیر است؛ داده‌هایی که علاوه بر نمونه‌های ورودی و برچسب‌های خروجی، حاوی اطلاعات صریح درباره‌ی منطق تصمیم‌گیری مدل، نواحی معنادار تصویر و شواهد بصری مرتبط با هر پیش‌بینی هستند. این نوع فراداده‌ها نقش یک لایه‌ی ارتباطی[۶] میان انسان و مدل را ایفا می‌کنند و امکان انتقال دانش سطح بالا درباره‌ی «چرایی» و «چگونگی» تصمیم‌گیری را فراهم می‌سازند [۳]. در نتیجه، مدل نه‌تنها قادر به انجام وظیفه‌ی پیش‌بینی خواهد بود، بلکه می‌تواند مسیر استدلال خود را نیز به‌صورت شفاف و قابل تحلیل ارائه دهد. از سوی دیگر، در حوزه‌ی داده‌های تصویری، فقدان مجموعه‌داده‌هایی با حقیقت مطلق توضیح‌پذیری، بار اصلی تولید این اطلاعات را بر دوش نیروی انسانی و زیرساخت‌های محاسباتی[۷] قرار می‌دهد. این امر مستلزم افزودن حاشیه‌نویسی‌های دستی نظیر مرزبندی دقیق اشیا، ناحیه‌بندی بخش‌های کلیدی تصویر یا تعیین اهمیت پیکسلی نواحی مختلف در تصمیم مدل است. تولید چنین داده‌هایی معمولاً زمان‌بر و پرهزینه بوده و نیازمند دانش تخصصی و

[۱] Post-hoc
[۲] Intrinsic
[۳] Explanation-Guided Learning (EGL)
[۴] Intelligence
[۵] ground truth
[۶] Communications
[۷] Computers

[Downloaded from ic4i-journal.ir on 2026-11-08]

بررسی دقیق هر نمونه به‌صورت مجزا است. در نتیجه، نبود یک چارچوب منسجم در EGL، به‌عنوان یک گلوگاه اساسی در توسعه‌ی مدل‌های شفاف، قابل اعتماد و قابل کنترل مطرح می‌شود.

۱-۱. مقدمه و هدف

در این پژوهش با بهره‌گیری از EGL، راهکاری به منظور افزایش تمرکز مدل‌های طبقه‌بندی نسبت به سوژه اصلی و استخراج ویژگی‌های کلیدی و قابل تمایز از تصویر ارائه شده است. به همین منظور برای تولید مجموعه داده شامل حقیقت مطلق توضیح‌پذیری و کاهش هزینه‌ی زمانی و انسانی، روش ارائه شده مبتنی بر خودکارسازی تولید فراداده‌های توضیح‌پذیر ارائه شده است. در گام نخست، از مدل‌های از پیش آموزش‌دیده و قابل اعتماد در حوزه‌ی بینایی ماشین، به‌ویژه در وظایفی مانند بخش‌بندی تصویر[۱]، استفاده شده است تا به‌صورت خودکار ماسک‌های دودویی مرتبط با اشیای اصلی در هر تصویر تولید شوند. این ماسک‌ها نشان‌دهنده‌ی نواحی کلیدی در تصویر هستند که بیشترین تأثیر را در تصمیم مدل دارند. سپس مجموعه‌داده‌ای ترکیبی شامل تصویر اصلی، ماسک دودویی استخراج‌شده، و برچسب نهایی تصویر ایجاد گردید. این ترکیب، نوعی حقیقت مطلق تقریبی برای توضیح‌پذیری ایجاد می‌کند که با دقت بالایی به واقعیت نزدیک است. در ادامه، این داده‌های ترکیبی به عنوان ورودی برای آموزش شبکه‌های عصبی با رویکرد یادگیری مبتنی بر توضیح‌پذیری مورد استفاده قرار گرفتند تا مدل نه تنها در وظیفه‌ی طبقه‌بندی عملکرد بهتری داشته باشد، بلکه قابلیت تبیین تصمیمات خود را نیز در سطح پیکسلی یا ناحیه‌ای ارتقا دهد.

در این پژوهش، این سؤال مطرح شده است که آیا می‌توان با ترکیب قابلیت‌های مدل **SAM۲** در تولید خودکار ماسک‌های توضیح‌پذیری و طراحی یک تابع ضرر مبتنی بر تمرکز، مدل‌های طبقه‌بندی را به یادگیری ویژگی‌هایی هدایت کرد که توجه بیشتری به سوژه اصلی تصویر داشته باشند، بدون آنکه عملکرد طبقه‌بندی کاهش یابد؟

برای پاسخ به این سؤال، روشی مبتنی بر ترکیب **SAM۲** و یک تابع ضرر جدید (**FS Loss**) ارائه شده است که با استفاده از داده‌های توضیح‌پذیری تولیدشده به‌صورت خودکار، مدل را در فرآیند آموزش به تمرکز بیشتر بر نواحی مرتبط با سوژه اصلی هدایت می‌کند. نتایج نشان می‌دهد که روش پیشنهادی ضمن افزایش تمرکز مدل‌های طبقه‌بندی بر سوژه اصلی و استخراج ویژگی‌های کلیدی و متمایز، وابستگی به مداخله انسانی در تولید داده‌های توضیح‌پذیری را نیز کاهش می‌دهد. این ویژگی، امکان تولید مجموعه‌داده‌های استاندارد، مقیاس‌پذیر و قابل اتکا برای ارزیابی کمی روش‌های توضیح‌پذیری را فراهم می‌سازد.

در مجموع، نوآوری اصلی این پژوهش تابع ضرر پیشنهادی **FS Loss** به‌منظور هدایت مدل به یادگیری متمرکزتر است. این پژوهش گامی در جهت توسعه مدل‌های طبقه‌بندی با توضیح‌پذیری ذاتی بیشتر، شفافیت بالاتر و قابلیت اعتماد بیشتر در کاربردهای مبتنی بر یادگیری عمیق محسوب می‌شود.

روش پیشنهادی و نتایج به دست آمده در این پژوهش در ۵ بخش شرح داده شده است که به ترتیب شامل موارد زیر است:

- پژوهش‌های پیشین: در این بخش مروری بر کارهای پژوهشی پیشین صورت گرفته است.
- روش پیشنهادی: در این بخش ایده و نوآوری اصلی روش ارائه شده برای دستیابی به هدف افزایش تمرکز مدل بر روی سوژه اصلی شرح داده شده است.
- آزمایشات: در این فصل آزمایشات انجام گرفته برای روش پیشنهادی مطرح شده اند.
- نتایج: در این بخش نتایج آزمایشات انجام شده ذکر شده است.
- بحث و نتیجه‌گیری: در این بخش نتیجه‌گیری نهایی از روش ارائه شده و تحلیل و تفسیر نتایج به دست آمده قرار داده شده است.

۲-۱. پژوهش‌های پیشین

روش آموزش همراه با راهنمایی‌های توضیح‌پذیری

[۱] Image Segmentation

[Downloaded from ic4i-journal.ir on 2026-11-08]

هدف EGL بهینه‌سازی همزمان عملکرد پیش‌بینی مدل و توضیحات متناظر آن است. به طور معمول، روش‌های EGL از تابع هدف زیر برای آموزش مدل $\boldsymbol{f}$ استفاده می‌کنند:

$$\boldsymbol{min_f\, E_{(x,y)}}\Big[\underbrace{\mathcal{L}_{\mathbf{cls}}(\boldsymbol{f}(\boldsymbol{x}),\boldsymbol{y})}_{\textit{task supervision}} + \underbrace{\boldsymbol{\alpha}\mathcal{L}_{\mathbf{exp}}\,(\boldsymbol{g}(\boldsymbol{f},\langle \boldsymbol{x},\boldsymbol{y}\rangle),\boldsymbol{M})}_{\textit{explanation supervision}} + \underbrace{\boldsymbol{\beta\Omega}(\boldsymbol{g}(\boldsymbol{f},\langle \boldsymbol{x},\boldsymbol{y}\rangle))}_{\textit{explanation regularization}}\Big]$$

که در آن $\boldsymbol{g}(\boldsymbol{f},\langle \boldsymbol{x},\boldsymbol{y}\rangle)$ توضیح مدل $\boldsymbol{f}$ در ارتباط با نمونه $\boldsymbol{x}$ و برچسب $\boldsymbol{y}$ است. $\boldsymbol{M}$ نشان‌دهنده‌ی «توضیح درست» می‌باشد که معمولاً توسط حاشیه‌نویسی انسانی تعیین می‌شود [۱]. این تابع هدف شامل سه جزء کلیدی است:

- **تابع ضرر نظارت بر وظیفه**[۱]: مانند آنتروپی متقاطع در وظایف طبقه‌بندی که مدل را به سمت پیش‌بینی‌های دقیق‌تر هدایت می‌کند.
- **تابع ضرر نظارت بر توضیح**[۲]: که دانش پیشینی صریح را وارد فرآیند می‌کند و تضمین می‌نماید توضیحات مدل با آنچه یک توضیح صحیح باید باشد هم‌راستا باشد.
- **تابع ضرر متعادل‌سازی توضیح**[۳]: که ویژگی‌های مطلوب برای توضیحات را تحمیل می‌کند، مانند ترویج تُنُکی (sparsity) و متمرکز کردن توجه بر مرتبط‌ترین ویژگی‌ها.

روش FocusEx

این روش یک چارچوب خودنظارتی EGL است که با کاهش حواس‌پرتی ناشی از نویز پس‌زمینه، مدل را بر اهداف اصلی متمرکز می‌کند، بدون آنکه نیاز به برچسب‌گذاری اضافی باشد. FocusEx به‌طور یکپارچه با مدل‌های از پیش آموزش‌دیده‌ی در مقیاس بزرگ[۴] ادغام می‌شود. این مدل‌ها با مشخص کردن ویژگی‌های کلیدی، نوعی نظارت ضعیف فراهم می‌کنند. نواحی برجسته‌شده به‌عنوان مناطق مورد توجه، مدل را هدایت می‌کنند تا بر اهداف اطلاعاتی تمرکز کند و در عین حال پس‌زمینه نامرتبط را با محدودسازی نقشه‌های توضیح به این نواحی استخراج‌شده، نادیده بگیرد. در این پژوهش، با مقایسه نقشه حرارتی[۵] با ناحیه هدف، تابع ضرر تقویت‌شده با توضیح، سازوکار توجه مدل را تنظیم می‌کند و باعث می‌شود مدل در طول آموزش به‌طور دقیق روی ناحیه هدف متمرکز شود و وابستگی به پس‌زمینه کاهش یابد. این امر قابلیت توضیح‌پذیری و تعمیم‌پذیری مدل را بهبود می‌بخشد. سازوکار آموزشی پیشنهادی، به‌طور مؤثر مدل را در جهت تمرکز بهتر بر نواحی حیاتی و در عین حال بهینه‌سازی دقت پیش‌بینی هدایت می‌کند و در نهایت توضیحات قابل‌اعتمادتر تولید می‌نماید [٤]. توابع ضرر به کار گرفته شده در این پژوهش به این صورت هستند:

- **تابع ضرر طبقه‌بندی:** برای این منظور از تابع Cross Entropy استفاده شده است.
- **تابع ضرر توضیح‌پذیری:** برای این منظور از تفاوت نقشه حرارتی تولید شده با ماسک تصویر با استفاده از Cross Entropy استفاده شده است.
- **تابع ضرر متعادل‌سازی پراکندگی (Sparsity) و داده افزایی:** برای متعادل‌سازی پراکندگی از تابع $L_۲$ در نقشه حرارتی استفاده شده است. و برای جلوگیری از بیش‌برازش از تابع ضرر (mixup) با راهبرد تولید نمونه‌های درون‌یابی‌شده[۶] بین جفت ورودی‌ها استفاده شده است.

با جلوگیری از بیش‌برازش به الگوهای کاذب، FocusEx در آزمایش‌ها عملکرد بهتری نسبت به روش‌های پیشرفته‌ی موجود از نظر دقت پیش‌بینی و تعمیم‌پذیری در چندین مجموعه‌داده‌ی معیار نشان می‌دهد [٤]. اما از نظر کیفیت توضیحات ارائه شده و نقشه‌های حرارتی تولید شده، روش FocusEx در افزایش تمرکز شبکه عصبی با مشکل مواجه است و تصاویر نقشه حرارتی‌های تولید شده با این روش تمرکز مناسبی نسبت به تمام

[۱] Task supervision loss
[۲] Explanation supervision loss
[۳] Explanation regularization loss
[۴] Pretrained Models
[۵] heatmap
[۶] interpolated samples

[Downloaded from ic4i-journal.ir on 2026-11-08]

سوژه اصلی نداشته است. عدم تمرکز مدل به تمام ویژگی‌های سوژه اصلی موجب کاهش دقت مدل در مجموعه داده‌هایی می‌شود که دارای فضای ویژگی مشابهی هستند. با افزایش تمرکز شبکه عصبی بر سوژه اصلی مدل توانایی درک بهتر جزئیات سوژه و عملکرد بهتری در انجام وظیفه اصلی طبقه‌بندی دارد.

در پژوهش [۵]، چارچوبی با نام یادگیری هدایت‌شده با تبیین ترکیبی (H-EGL) پیشنهاد شده است که محدودیت‌های خودنظارتی و هدایت‌شده توسط انسان را ترکیب می‌کند تا هم‌ترازی مکانیزم توجه را بهبود داده و قابلیت تعمیم مدل را افزایش دهد. بخش خودنظارتی H-EGL از توجه متمایزکننده‌ی کلاسی بدون تکیه بر پیش‌فرض‌های محدودکننده بهره می‌برد، که موجب افزایش پایداری و انعطاف‌پذیری می‌شود [۵]. این روش بر روی طبقه‌بندی تصاویر رادیوگرافی قفسه سینه با استفاده از Vision Transformer (ViT) ارزیابی شده است. نتایج نشان می‌دهد که H-EGL نسبت به دو روش پیشرفته‌ی موجود در حوزه‌ی یادگیری هدایت‌شده با تبیین (EGL) عملکرد بهتری دارد و دقت طبقه‌بندی و قابلیت تعمیم بالاتری ارائه می‌دهد. علاوه بر این، نقشه‌های توجه تولیدشده توسط این روش هم‌ترازی بهتری با تخصص انسانی نشان می‌دهند [۵].

H-EGL یک رویکرد ترکیبی از نوع یادگیری هدایت‌شده با تبیین (EGL) است که برای ترنسفورمرهای بینایی (ViT) طراحی شده است. این مدل ورودی‌های تصویری (مانند تصاویر رادیوگرافی قفسه سینه) و اطلاعات متنی (مانند برچسب‌های بیماری نظیر آتلکتازی یا کاردیومگالی (را از طریق یک معماری رمزگذار–رمزگشا مبتنی بر ViT با هم ترکیب می‌کند. در این ساختار، رمزگذار متنی ثابت (فریز شده) بوده و اجزای ViT قابل آموزش هستند. مدل در نهایت پیش‌بینی‌های کلاسی و نقشه‌های توجه را تولید می‌کند.

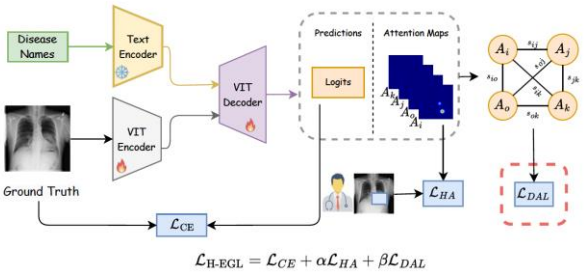
چارچوب H-EGL شامل دو مؤلفه‌ی کلیدی است:

- ماژول خودنظارتی یادگیری توجه متمایزکننده (DAL) که به تولید الگوهای توجه متمایز برای هر کلاس کمک می‌کند
- ماژول نظارت‌شده‌ی هم‌ترازی انسان–هوش‌مصنوعی[۱]، که تمرکز آن بر تطبیق نقشه‌های توجه با نشانه‌گذاری‌های متخصصان (مانند جعبه‌های محصورکننده[۲]) است [۵].

تابع ضرر کلی آموزش شامل سه بخش اصلی است:

- آنتروپی متقاطع ($\mathcal{L}_{CE}$) برای انجام طبقه‌بندی
- هم‌ترازی انسان–هوش‌مصنوعی ($\mathcal{L}_{HA}$) برای هم‌راستا کردن نقشه‌های توجه با نشانه‌های انسانی
- توجه متمایزکننده ($\mathcal{L}_{DAL}$) برای ایجاد جداسازی میان نقشه‌های توجه مربوط به کلاس‌های مختلف.

شکل ۱ مفهوم تمایز توجه[۳] را در یک سناریوی چهاربخشی نشان می‌دهد. این رویکرد برای هر تعداد کلاس قابل تعمیم است.



شکل ۱. چارچوب یادگیری هدایت‌شده با تبیین ترکیبی روش H-EGL [۵].

۱-۳. معیارهای ارزیابی توضیح‌پذیری

در رویکرد استفاده از راهنمایی‌های توضیح‌پذیری در آموزش مدل (EGL)، برای ارزیابی مدل‌های آموزش داده شده علاوه بر معیارهای پیش‌بینی مانند مساحت زیر منحنی (ACU) و دقت (Accuracy)، معیارهای زیر نیز مورد بررسی قرار می‌گیرند.

نسبت مساحت (AR)[۴]

این معیار همپوشانی بین نواحی با فعال‌سازی بالا در نقشه‌های Grad-CAM و ماسک شیء را اندازه‌گیری می‌کند [۴] که به‌صورت رسمی به شکل زیر تعریف می‌شود:

---

[۱] Human-AI Alignment

[۲] Bounding box

[۳] attention distinctiveness

[۴] Area Ratio

[Downloaded from ic4i-journal.ir on 2026-11-08]

$$AR(G,M)=\frac{\sum_{i,j}\left(I\left(G_{ij}(x)>\tau\right)\cdot M(i,j)\right)}{\sum_{i,j}I\left(G_{ij}(x)>\tau\right)}$$

که در آن $\boldsymbol{\tau}$ آستانه‌ی فعال‌سازی (به‌طور پیش‌فرض ۰٫۵) است و $\boldsymbol{M(i,j)}$ برچسب حقیقت مطلق **(ground truth)** ماسک هدف می‌باشد.

کامل‌بودن توضیح (Comp)[۱]

کیفیت توضیح را با اندازه‌گیری اختلاف احتمال پیش‌بینی بین ورودی اصلی ($\boldsymbol{x}$) و ورودی‌ای که بخش‌های پشتیبان آن حذف شده‌اند ($\boldsymbol{X_i \setminus g(f,\langle X_i,Y_i\rangle)}$)، می‌سنجد [۶], [۷], [۸], [۹], [۱۰]:

$$Comprehensiveness = f(X_i)_{Y_i} - f\left(X_i \setminus g(f,\langle X_i,Y_i\rangle)\right)_{Y_i}$$

که در آن $\boldsymbol{X_i}$ تصویر ورودی، $\boldsymbol{Y_i}$ برچسب (کلاس) تصویر ورودی، $\boldsymbol{f(X_i)_{Y_i}}$ پیش‌بینی مدل برای تصویر $\boldsymbol{X_i}$ و برچسب $\boldsymbol{Y_i}$ ، $\boldsymbol{g(f,\langle X_i,Y_i\rangle)}$ نقشه حرارتی تولید شده برای تصویر $\boldsymbol{X_i}$ ، برچسب $\boldsymbol{Y_i}$ با استفاده از مدل $\boldsymbol{f(.)}$ و عملگر \ به‌معنای حذف بخش‌های پشتیبان یا استدلال‌های توضیحی $\boldsymbol{g(f,\langle X_i,Y_i\rangle)}$ از ورودی $\boldsymbol{X_i}$ است.

مقادیر بیشتر معیار Comp نشان می‌دهد که ویژگی‌های استخراج‌شده بهتر می‌توانند پیش‌بینی مدل را پشتیبانی کنند.

کفایت توضیح (Suff)[۲]

این معیار بررسی می‌کند که آیا ویژگی‌های استخراج‌شده به‌قدر کافی برای پیش‌بینی دقیق مدل کافی هستند یا خیر. این معیار به این صورت تعریف می‌شود [۷], [۹], [۱۱], [۱۲]:

$$Suff = f_y(x) - f_y\left(g(f,\langle x,y\rangle)\right)$$

مقادیر کمتر **Suff** نشان می‌دهد که ویژگی‌های استخراج‌شده بهتر می‌توانند پیش‌بینی مدل را پشتیبانی کنند.

۱-۴. روش تحقیق یا اصول و تئوری مقاله

فرآیند آموزش و ارزیابی مدل به‌صورت ساختاریافته در قالب یک چرخه یادگیری خودکار پیاده‌سازی شده است. در ابتدا، مدل ResNet۱۸ با وزن‌های از پیش‌آموزش‌دیده روی مجموعه داده ImageNet بارگذاری می‌شود. سپس با استفاده از مدل SAM[۳] برای تصاویر موجود در مجموعه داده ماسک‌های دودویی تولید می‌شود. در فرآیند تولید ماسک‌های دودویی برای تعیین نقطه دستور (Prompt)، به این صورت عمل می‌شود، از یک شبکه ResNet۵۰ از پیش آموزش داده شده برای تولید نقشه حرارتی Grad-CAM برای تصاویر استفاده می‌شود. سپس نقطه‌ای که نقشه حرارتی Grad-CAM تصویر در آن بیشنیه است پیدا می‌شود و آن نقطه به عنوان دستور به SAM۲ ارسال می‌شود تا فرآیند بخش‌بندی دودویی نسبت به آن نقطه صورت گیرد. در مرحله بعد تابع ضرر FS Loss با اهداف زیر تعریف می‌شوند:

- طبقه‌بندی درست تصاویر
- افزایش تمرکز شبکه عصبی بر نواحی معنادار و سوژه اصلی تصویر
- کاهش نقاط حداکثری و پراکندگی در نقشه حرارتی

این تابع با بهره‌گیری از نقشه‌های حرارتی حاصل از Grad-CAM، تراکم پاسخ‌های مدل را در نواحی خاصی از تصویر تنظیم کرده و از توجه مدل به بخش‌های غیرمرتبط جلوگیری می‌کند. به‌منظور بهینه‌سازی پارامترهای مدل، از بهینه‌ساز Adam با نرخ یادگیری $\boldsymbol{10^{-4}}$ و جمله متعادل‌ساز وزن $\boldsymbol{L_2}$ برابر با ($\boldsymbol{10^{-4}}$) استفاده شده است. نمای کلی شبکه طراحی شده در شکل ۲ نشان داده شده است.

فرآیند آموزش در طی ۱۵ دوره انجام می‌شود. پس از اتمام هر دوره آموزشی، مدل در حالت ارزیابی قرار گرفته و روی داده‌های اعتبارسنجی آزمایش می‌شود. در این مرحله، مدل پیش‌بینی‌های خود را بدون به‌روزرسانی وزن‌ها انجام داده و مقادیر زیان و دقت محاسبه می‌شوند تا عملکرد آن روی داده‌های مشاهده نشده بررسی گردد. بدین‌ترتیب، روند کاهش زیان و افزایش دقت در مجموعه‌های آموزش و اعتبارسنجی در طول دوره‌های مختلف قابل ردیابی بوده و معیارهایی مانند

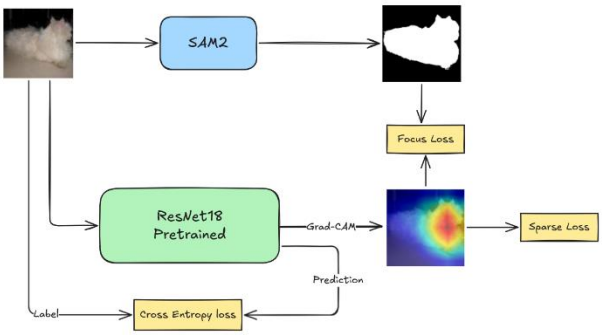
[۱] Comprehensiveness

[۲] Sufficiency

[۳] Segment Anything Model ۲

[Downloaded from ic4i-journal.ir on 2026-11-08]

Train/Test Loss و Train/Test Accuracy به عنوان شاخص‌های عملکردی مدل ذخیره می‌شوند.



$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \gamma_{sparse}\mathcal{L}_{Sparse} + \beta_{focus}\mathcal{L}_{focus}$$

شکل ۲. معماری شبکه عصبی پیشنهادی به همراه تابع ضرر طراحی شده.

در مجموع، این ساختار امکان آموزش یک مدل بینایی مبتنی بر ResNet۱۸ را با افزودن راهنمایی‌های توضیح‌پذیری از طریق مکانیزم‌های گرادیان‌محور را فراهم می‌کند. ترکیب استفاده از تابع ضرر طبقه‌بندی و راهنمایی‌های نقشه حرارتی Grad-CAM موجب می‌شود که، مدل در کنار دقت بالای طبقه‌بندی، آموزش ببیند تا به قسمت‌های بامعناتر تصویر و سوژه اصلی توجه کند و توجه به سایر قسمت‌های تصویر شامل پس‌زمینه کاهش یابد.

تابع ضرر FS Loss

برای بهره‌گیری از راهنمایی‌های توضیح‌پذیری در حین آموزش مدل، تابع ضرر ترکیبی که هدف آن، یادگیری ویژگی‌های مناسب افزایش دقت طبقه‌بندی، افزایش تمرکز بر سوژه اصلی و کاهش پراکندگی در نقشه‌های حرارتی است طراحی شده است. تابع ضرر FS Loss از بخش‌های زیر طراحی شده و شامل سه مؤلفه‌ی اصلی است:

- ضرر طبقه‌بندی: از تابع Cross Entropy استفاده شده است و علامت اختصاری آن ($\mathcal{L}_{CE}$) است.
- ضرر تمرکز: با علامت اختصاری ($\mathcal{L}_{Focus}$) است و زیان مرتبط با کیفیت تمرکز نقشه حرارتی بر سوژه اصلی تصویر را محاسبه می‌کند.
- ضرر پراکندگی: با علامت اختصاری ($\mathcal{L}_{Sparsity}$) است و مدل را مجبور می‌کند تا پراکندگی نقشه حرارتی‌های تولید شده را تا حد ممکن کم کند و نقشه حرارتی تُنُک از تصویر ورودی ایجاد کند.

ایده‌ی اصلی این تابع آن است که مدل علاوه بر یادگیری صحیح در سطح برچسب، در سطح ویژگی‌ها نیز رفتار قابل توضیح و متمرکز بر نواحی مرتبط با شیء هدف داشته باشد. این ساختار به مدل اجازه می‌دهد تا همزمان با حفظ عملکرد پیش‌بینی، ویژگی‌های درونی خود را از نظر بصری بهبود دهد و از تولید نقشه‌های غیرمتمرکز یا پر نویز جلوگیری کند.

زیان طبقه‌بندی[۱]

مولفه‌ی اول تابع ضرر، همان زیان استاندارد Cross-Entropy است که برای یادگیری درست برچسب‌ها به کار می‌رود:

$$\mathcal{L}_{CE} = -\frac{1}{N}\sum_{i=1}^{N} \log p\,(y_i|x_i)$$

که $\boldsymbol{p(y_i\,|x_i))}$ خروجی شبکه برای کلاس صحیح است. این بخش مسئول اصلی یادگیری وظیفه‌ی طبقه‌بندی است.

زیان تمرکز[۲]

برای اجبار مدل به یادگیری شباهت بین نقشه حرارتی تولیدشده و ماسک هدف، از معیار میانگین مربعات خطا (MSE) استفاده شده است:

$$\mathcal{L}_{focus} = \sum_{j=1}^{W}\sum_{i=1}^{H}\frac{1}{HW}\left(H_{pred}(i,\,j)\; - H_{mask}(i,j)\right)^2$$

در این رابطه، $\boldsymbol{H_{mask}}$ ماسک یا نقشه هدف است و $\boldsymbol{H_{pred}}$ نقشه حرارتی تولیدی مدل است. این زیان، شبکه را وادار می‌کند تا در ناحیه‌های مشخص‌شده توسط ماسک تمرکز کند و ویژگی‌های محدود شده به ماسک هر تصویر را در نقشه حرارتی انتخاب کند.

زیان پراکندگی[۳]

[۱] Classification Loss

[۲] Focus Loss

[۳] Sparsity Loss

[Downloaded from ic4i-journal.ir on 2026-11-08]

برای جلوگیری از تولید نقشه‌های حرارتی با نواحی فعال گسترده و پر نویز، یک جمله پراکندگی بر پایه‌ی نرم (L_1) تعریف شده است [٤]:

$$\mathcal{L}_{sparse} = \sum_{j=1}^{W}\sum_{i=1}^{H}\frac{1}{HW}\left|H_{pred}(i,j)\right|$$

این مؤلفه باعث می‌شود مدل فقط در نواحی واقعاً مهم فعال شود و سایر نواحی را خاموش نگه دارد. در نتیجه، نقشه حرارتی نهایی، فشرده‌تر و از نظر بصری قابل تفسیرتر خواهد بود.

در مجموع، تابع ضرر پیشنهادی را می‌توان به شکل زیر نوشت:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \gamma_{sparse}\mathcal{L}_{Sparse} + \beta_{focus}\mathcal{L}_{focus}$$

این تابع ضرر سه هدف کلیدی را دنبال می‌کند:

- حفظ دقت مدل در سطح پیش‌بینی با استفاده از قرار دادن Cross-Entropy در تابع ضرر.
- افزایش هم‌راستایی بین نواحی تمرکز مدل و نواحی واقعی شیء با قرار دادن Focus Loss.
- بهبود فشردگی و تفسیرپذیری نقشه‌های حرارتی با استفاده از Sparsity Loss.

به این ترتیب، مدل نه تنها عملکرد طبقه‌بندی مناسبی خواهد داشت، بلکه از نظر تبیین رفتار نیز قابل اعتمادتر و هم‌راستا با دانش انسانی خواهد بود.

فرآیند تولید ماسک دودویی

در چارچوب کار حاضر، از ماژول SAM۲ برای تولید ماسک دودویی استفاده شده تا جداسازی دقیق پس‌زمینه و پیش‌زمینه روی تصاویر فراهم گردد. مدل SAM۲ در این پژوهش به‌عنوان یک مدل بنیادین برای تقسیم‌بندی قابل‌پرسش[1] به‌کار گرفته شده‌است. طبق توصیف رسمی SAM۲، این مدل گامی به‌سوی حل مسئله تقسیم‌بندی تصویر و ویدئو با قابلیت پاسخ به نشانگرهای مختلف است. در این نسخه، تصاویر همانند یک ویدئوی تک‌فریمی در نظر گرفته شده‌اند و از یک معماری مبتنی بر ترنسفورمر با حافظه جریان‌پذیر برای پردازش در زمان‌واقعی بهره‌گیری شده است. علاوه بر ساختار مدل، یک موتور داده‌سازی با تعامل در حلقه مدل[2] توسعه داده شد که با کمک تعامل انسانی مجموعه‌داده‌های ویدیویی-تقسیمی را گردآوری و بهبود می‌بخشد. نتیجه این فرآیند، آموزش SAM۲ بر روی داده‌های جمع‌آوری‌شده و تقویت عملکرد آن در حوزه‌های بصری متنوع بوده است [۲].

روند پردازش به‌صورت گام‌به‌گام به صورت زیر طراحی شده‌است:

- ابتدا برای هر تصویر یک نقشه حرارتی تولید شده که نشان‌دهنده اهمیت یا فعال‌سازی مقاطع مختلف تصویر برای معیار موردنظر است.
- سپس مکان نقطه‌ای که مقدار نقشه گرمایی در آن حداکثر بود (نقطه ماکزیمم) با استفاده از مدل از پیش آموزش داده شده ResNet۵۰ با وزن‌های ImageNet استخراج شده.
- این مختصات نقطه به‌عنوان ورودی دستور[3] در اختیار مدل SAM۲ قرار داده شده تا بر اساس آن ناحیه مربوط به شیء موردنظر از زمینه تفکیک شود.
- پس از تولید ماسک دودیی توسط مدل SAM۲ تمام تصاویر تولید شده توسط ناظر انسانی بررسی می‌شوند و در صورت وجود خطا در فرایند تولید ماسک از مجموعه داده نهایی کنار گذاشته می‌شوند.

با استفاده از نقطه تعیین‌شده، SAM۲ یک ماسک باینری برای تصویر تولید نموده که به‌وضوح نواحی پیش‌زمینه را از پس‌زمینه جدا می‌کند. تولید این ماسک منجر به تعریف توابع ضرری برای آموزش مدل می‌شود که تنها بر روی بخش‌های مرتبط تصویر فرآیند یا محاسباتی را انجام می‌دهند و اثر نویز و اطلاعات نامربوط را کاهش می‌دهند. همچنین استفاده از نقطه ماکزیمم نقشه حرارتی به‌عنوان نقطه راهنما، فرآیند تولید ماسک را هدفمند و کم‌هزینه می‌کند، زیرا فرآیند تولید ماسک از تصاویر را از یک فرآیند کاملا دستی تبدیل به یک فرآیند کاملا خودکار می‌کند که به‌جای ارائه چندین نشانگر یا جستجوی گسترده

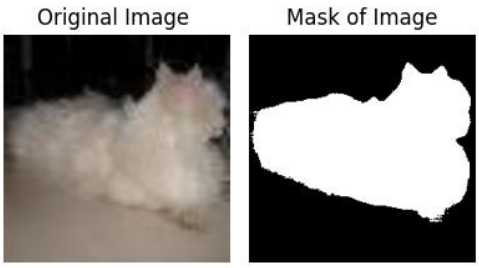
[1] promptable segmentation

[2] model-in-the-loop data engine

[3] prompt

[Downloaded from ic4i-journal.ir on 2026-11-08]

توسط عامل انسانی، تنها یک نقطه کلیدی که اتفاقا مدل توجه بیشتری به آن داشته است را انتخاب کرده تا مدل بتواند مرزهای شیء را بازشناسی کند. نمونه مرزبندی تصاویر در شکل ۳ نشان داده شده‌است.



شکل ۳. نمونه مرزبندی تصاویر مجموعه داده با استفاده از مدل SAM۲.

ادغام نقشه حرارتی و SAM۲ مزایای عملی متعددی دارد. اولاً، تمرکز پردازش بر بخش‌های معنادار تصویر افزایش یافته و در نتیجه تحلیل‌های آماری و بصری قابل‌اطمینان‌تر می‌شوند. ثانیاً، روش پیشنهادی به‌صورت عمومی‌تری قابل تعمیم است، زیرا به‌کارگیری یک نقطه ماکزیمم از نقشه حرارتی می‌تواند برای انواع مختلف تصاویر و سناریوها به‌عنوان یک نشانه ساده اما مؤثر عمل نماید.

در نهایت، بهره‌گیری از SAM۲ که با داده‌های بزرگ و متنوع آموزش دیده است، سبب می‌شود تا تفکیک مرزهای پیش‌زمینه با دقت و پایداری بیشتری انجام شود و نیاز به تنظیمات دستی یا طراحی ماسک‌های تخصصی برای هر تصویر تا حد زیادی کاهش یابد.

۲. یافته‌ها

۲-۱. مجموعه داده انتخاب شده

در این پژوهش از مجموعه‌داده Tiny ImageNet برای انجام آزمایش‌ها استفاده شده است. این مجموعه‌داده نسخه‌ای فشرده و کوچک‌شده از ImageNet است و به‌صورت گسترده در ارزیابی مدل‌های یادگیری‌عمیق مورد استفاده قرار می‌گیرد [۱۴]. مشخصات این مجموعه‌داده به‌صورت زیر است:

- شامل ۱۰۰،۰۰۰ تصویر در ۲۰۰ کلاس مختلف (هر کلاس دارای ۵۰۰ نمونه تصویری) است.
- تمامی تصاویر به اندازه‌ی ٦٤×٦٤ پیکسل تغییر مقیاس یافته‌اند.
- برای هر کلاس، ۵۰۰ تصویر برای آموزش، ۵۰ تصویر برای اعتبارسنجی و ۵۰ تصویر برای آزمون در دسترس است.

به‌منظور کاهش هزینه‌ی محاسباتی و بررسی دقیق‌تر نتایج آزمایشات، تنها ۱۰ کلاس از میان ۲۰۰ کلاس انتخاب شده است. در انتخاب کلاس‌ها سعی شده است موضوع کلاس‌ها متمایز از یکدیگر باشند تا بتوان بررسی دقیق‌تری بر روی نتایج مدل‌های آموزش داده شده با این مجموعه داده انجام داد. تصاویر انتخاب‌شده، طی فرآیند پیش‌پردازش به اندازه‌ی ٢٢٤×٢٢٤ پیکسل تغییر اندازه داده شده‌اند.

کلاس‌های انتخاب‌شده در جدول ۱ آمده‌اند.

جدول ۱. برچسب‌های انتخاب شده از مجموعه داده TinyImageNet [۱۱].

نام فارسی کلاس	عنوان انگلیسی
صفحه کلید کامپیوتر	computer_keyboard
میز	desk
ماهیتابه	frying_pan
باجه تلفن عمومی	pay-phone
عینک آفتابی	sunglasses
پرتقال	orange
پیتزا	pizza
اسپرسو	espresso
گربه ایرانی	Persian_cat
زنبور عسل	bee

همچنین برای مقایسه روش پیشنهادی با سایر پژوهش‌های پیشین، روش پیشنهادی بر روی مجموعه داده VLCS نیز پیاده‌سازی شده است و سپس نتایج به دست آمده با نتایج گزارش شده در روش FocusEx مقایسه شده‌اند.

۲-۲. آزمایشات انجام شده

در این پژوهش برای آزمایش و ارزیابی روش پیشنهادی، شبکه ResNet۱۸ (همراه با وزن‌های از پیش آموزش داده شده روی مجموعه داده ImageNet) به عنوان مدل پایه انتخاب شده و لایه تماما متصل ۱۰۰۰ کلاسه آن با یک لایه تماما متصل ۱۰ کلاسه جایگزین شده است. با توجه به نوع تابع ضرر تعریف شده و نحوه استفاده از آن در حین آموزش برای عملکرد مناسب

[Downloaded from ic4i-journal.ir on 2026-11-08]

و قابل توجه، بهتر است از زیان‌های $\mathcal{L}_{focus}$ و $\mathcal{L}_{sparse}$ در زمانی استفاده شود که مدل در حال آموزش ویژگی‌های اصلی مسئله بوده و فضای عمومی مسئله را قبلا آموزش دیده باشد. بدین ترتیب محل استفاده صحیح توابع ضرر پیشنهاد شده در این حالت‌ها می‌باشد:

- استفاده از $\mathcal{L}_{focus}$ و $\mathcal{L}_{sparse}$ پس از چند دوره آموزش با استفاده از تابع ضرر طبقه‌بندی.
- استفاده از $\mathcal{L}_{focus}$ و $\mathcal{L}_{sparse}$ برای بهینه سازی مدل از پیش آموزش داده شده بر روی مسئله جدید.

آزمایشات صورت گرفته در این پژوهش طبق حالت دوم صورت گرفته است. شبکه ResNet۱۸ با فراپارامترهای $\boldsymbol{\beta_{focus}}$ و $\boldsymbol{\gamma_{sparse}}$ متفاوت آموزش داده شده و نتایج آن در دو فاز طبقه‌بندی و توضیح‌پذیری مقایسه شده‌اند. همچنین برای مقایسه بهتر تاثیر روش پیشنهادی، به صورت جداگانه شبکه ResNet۱۸ تنها با تابع ضرر $\boldsymbol{\mathcal{L}_{CE}}$ نیز آموزش داده شده است تا مقایسه صحیحی بین مدل‌های آموزش داده شده با راهنمایی توضیح‌پذیری و مدل‌های آموزش داده شده با روش‌های معمول صورت گیرد.

۳. نتایج

۳-۱. مقایسه مدل‌های توضیح‌پذیر و غیر توضیح‌پذیر

در این بخش ابتدا مدل‌های آموزش داده شده همراه با فراپارامترهای آن بررسی می‌شوند و سپس به بررسی نتایج آن از جمله نمودارهای فرآیند آموزش مدل و مقایسه دقت هر مدل پرداخته می‌شود.

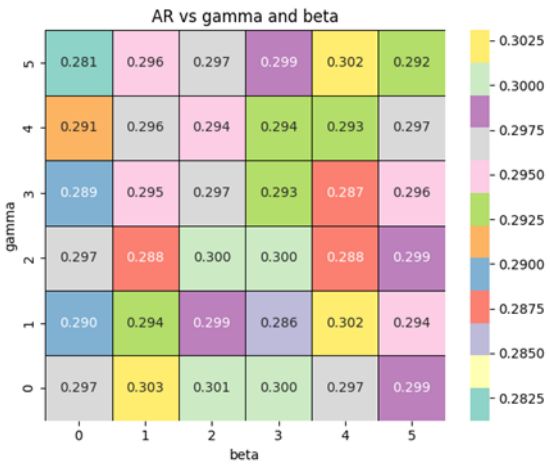
در این پژوهش شبکه عصبی **ResNet۱۸** با فراپارامترهای ثابت ذکر شده در جدول ۲ آموزش داده شده است:

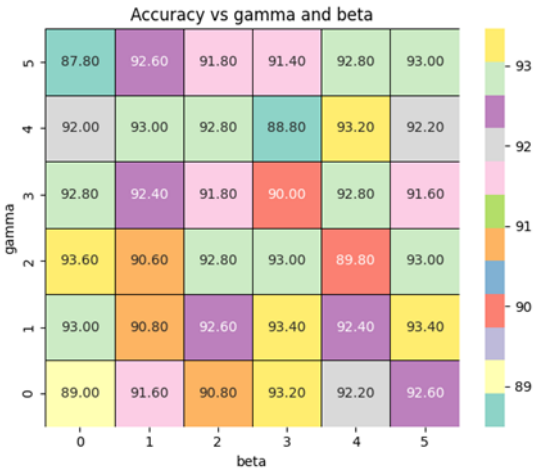
جدول ۲. فراپارامترهای تنظیم شده در آزمایشات صورت گرفته.

نوع فراپارامتر	مقدار
Weight decay	۰.۰۰۰۱
Learning rate	۰.۰۰۰۱
Batch size	۱۲۸
Grad-CAM target layer	آخرین لایه کانولوشن در مدل ResNet۱۸

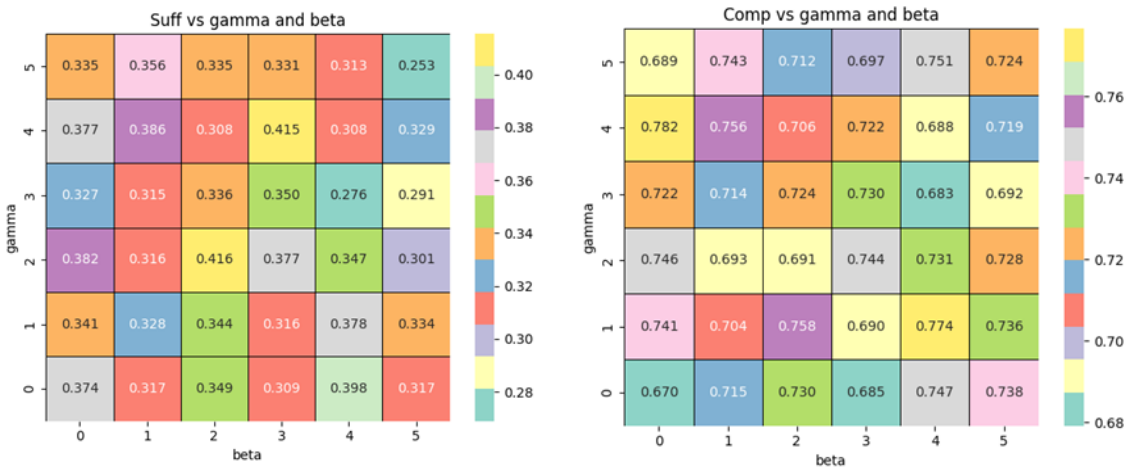
اما سایر فراپارامترها مانند $\boldsymbol{\beta_{focus}}$ و $\boldsymbol{\gamma_{sparse}}$ با مقادیر متفاوت آزمایش شده‌اند.

در ادامه نتایج آموزش مدل‌های در شکل ٤ قرار داده شده است.





[Downloaded from ic4i-journal.ir on 2026-11-08]



شکل ٤. نتایج معیارهای ارزیابی Suff, AR, Accuracy و Comp روش ارائه شده.

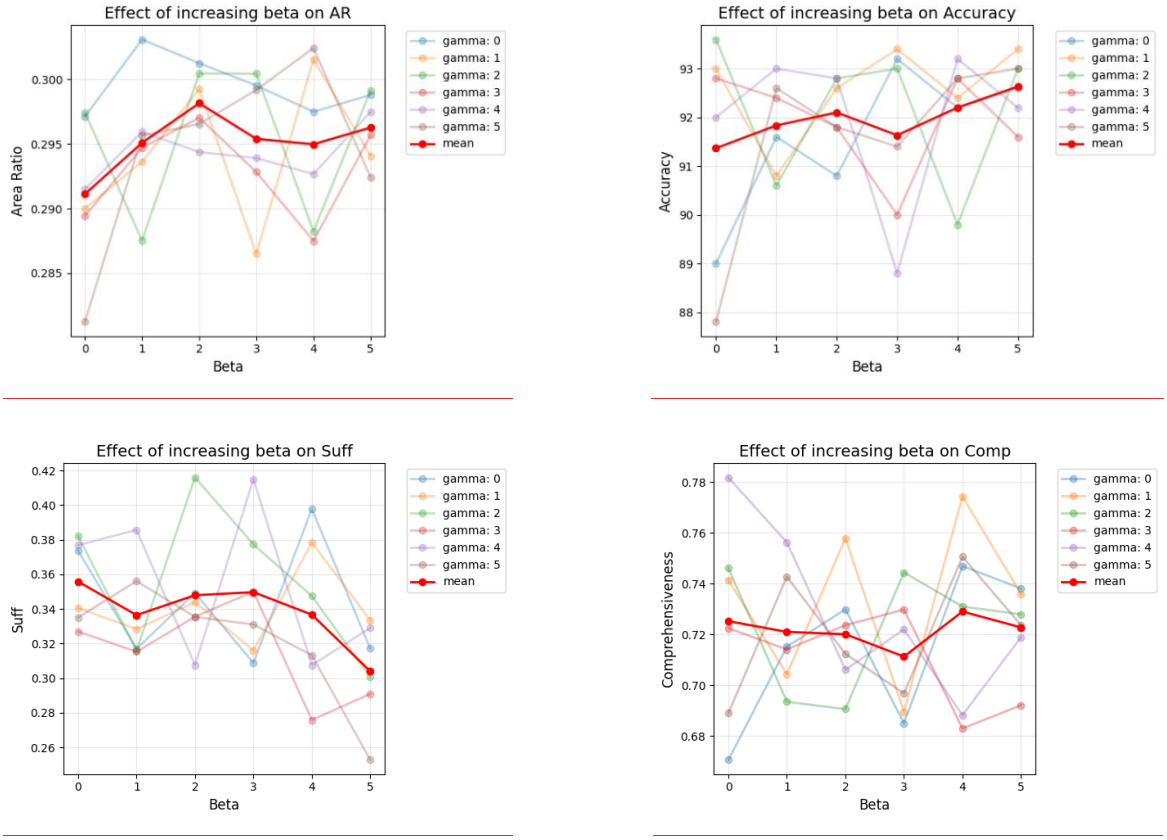
در شکل ٤ تمامی مدل‌ها در ۱۵ دوره آموزش داده شده و به مظور مقایسه بهتر سایر فراپارامترها ثابت شده‌اند. در نمودارها، محور $\boldsymbol{x}$ نشان دهنده مقادیر $\boldsymbol{\beta_{focus}}$ و محور $\boldsymbol{y}$ مقادیر $\boldsymbol{\gamma_{sparse}}$ آزمایش شده هستند. هر خانه از نمودارها بیانگر مقدار نهایی معیار ارزیابی مدل آموزش داده شده است. به عنوان مثال خانه پایین گوشه چپ، نشانگر مدل آموزش داده شده با فراپارامترهای $\boldsymbol{\gamma_{sparse} = 0.0}$ و $\boldsymbol{\beta_{focus} = 0.0}$ که یعنی تنها تابع ضرر موجود در این آزمایش خاص تابع ضرر $\boldsymbol{\mathcal{L}_{CE}}$ بوده است. اگر آزمایش با فراپارامترهای $\boldsymbol{\gamma_{sparse} = 0.0}$ و $\boldsymbol{\beta_{focus} = 0.0}$ را به عنوان مدل پایه در نظر بگیریم، با توجه به نتایج گزارش شده در تمامی معیارهای ارزیابی استفاده از تابع ضرر مبتنی بر توضیح‌پذیری از مدل پایه نتایج بهتری را کسب کرده است.

با بررسی دقیق‌تر می‌توان دریافت که، راهنمایی‌های توضیح‌پذیری به طور موثری در بهبود نتایج شبکه عصبی نقش داشته‌اند. در آزمایشاتی که صورت گرفته دقت نهایی مدل پایه بر روی مجموعه داده تست حدود ۸۹٪ است. اما این محدوده برای مدل‌هایی که از راهنمایی‌های توضیح‌پذیری استفاده می‌کنند در حدود ۹۲٪ $\pm$ ۱٪ است. که این مقدار در بیشترین حالت خود به ۹۳.۸٪ نیز می‌رسد. در همین قسمت با مقایسه نتیجه دقت مدل‌های آموزش داده شده می‌توان افزایش دقت پیش‌بینی برای مدل را برداشت کرد.

یکی از موارد کلیدی که به درک بهتر میزان تاثیر تابع ضرر طراحی شده کمک می‌کند، مقایسه نتایج معیارهای ارزیابی توضیح‌پذیری است. در شکل ٤ نتایج ارزیابی مدل‌های آموزش داده شده با معیارهای مطرح شده «در بخش معیارهای ارزیابی توضیح‌پذیری» نشان داده شده است. مدل‌های آموزش داده شده با روش ارائه شده در معیار AR نسبت به مدل پایه در بیشترین حالت حدود ۱٪ بهبود داشته‌اند. این میزان بهبود برای معیار Comp حدود ۷٪ و برای معیار Suff حدود ۱۰٪ است. با توجه به موارد بررسی شده، مدل‌های آموزش داده شده با روش پیشنهادی همزمان با برتری در دقت پیش‌بینی نسبت به مدل پایه، در ارزیابی‌های کمّی توضیح‌پذیری نیز برتری قابل توجهی دارند.

نتایج ارزیابی مدل‌های آموزش داده با معیارهای مطرح شده نشان داده شده است. مدل‌های آموزش داده شده با روش ارائه شده در معیار نسبت مساحت، نسبت به مدل پایه در بیشترین حالت حدود ۱٪ بهبود داشته‌اند. این میزان بهبود برای معیار کامل بودن توضیحات، حدود ۷٪ و برای معیار کفایت توضیحات، حدود ۱۰٪ است.

[Downloaded from ic4i-journal.ir on 2026-11-08]



شکل ۵. تاثیر تابع ضرر تمرکز و پارامترهای γ_{sparse} و β_{focus} بر دقت مدل و معیارهای کمی توضیح پذیری.

نتایج گزارش شده به صورت نمودار نشان داده شده‌اند. در این نمودارها، رفتار مدل در مواجهه با تغییرات شدت هر یک از مؤلفه‌های زیان به‌صورت نظام‌مند بررسی شده است تا حساسیت عملکرد مدل نسبت به این ابرپارامترها مشخص گردد. به‌طور مشخص، در نمودار مربوط به تأثیر بر دقت پیش‌بینی، پنج منحنی مجزا برای مقادیر مختلف γ_{sparse} ترسیم شده‌اند که به‌صورت خطوط کم‌رنگ نمایش داده شده‌اند. هر نقطه روی این منحنی‌ها بیانگر دقت نهایی مدل پس از آموزش با یک ترکیب مشخص از γ_{sparse} و β_{focus} است. این نمایش چندمنحنی امکان تحلیل هم‌زمان اثر متقابل دو پارامتر را فراهم می‌سازد. علاوه بر این، یک خط قرمز رنگ به‌عنوان روند میانگین تغییرات بر روی نمودار افزوده شده است تا تصویر کلان‌تری از رفتار سیستم ارائه دهد و وابستگی عمومی عملکرد مدل به تغییرات β_{focus} را آشکار سازد.

بر اساس روند نمایان‌شده توسط خط میانگین، می‌توان استنباط کرد که افزایش وزن مؤلفه تمرکز در تابع زیان (از طریق افزایش مقدار β_{focus}) به‌صورت کلی منجر به بهبود دقت پیش‌بینی شده است. این مشاهده نشان می‌دهد که هدایت شبکه به سمت تمرکز مکانی بیشتر بر نواحی معنادار تصویر، نه‌تنها باعث بهبود ساختار توضیحی نقشه‌های حرارتی می‌شود، بلکه از منظر عملکرد طبقه‌بندی نیز سودمند است. از سوی دیگر، پارامتر γ_{sparse} که نقش تنظیم شدت قید پراکندگی را بر عهده دارد، ابزاری برای پالایش دقیق‌تر ساختار نقشه‌های حرارتی محسوب می‌شود. تنظیم مناسب این پارامتر می‌تواند از ایجاد فعال‌سازی‌های گسترده و غیرهدفمند جلوگیری کرده و تمرکز ویژگی‌ها را به نواحی کلیدی محدود سازد. نتایج نشان می‌دهد که انتخاب مقدار بهینه برای γ_{sparse} علاوه بر ارتقای معیارهای

[Downloaded from ic4i-journal.ir on 2026-11-08]

توضیح‌پذیری (مانند تمرکز و تمایزپذیری نقشه‌ها)، می‌تواند موجب بهبود یا تثبیت دقت پیش‌بینی نیز شود.

در مجموع، تحلیل نمودارهای شکل ۵ نشان می‌دهد که ترکیب هم‌افزای دو مؤلفه تمرکز و پراکندگی در قالب ضرایب γ_{sparse} و β_{focus}، امکان تنظیم توازن میان دقت پیش‌بینی و کیفیت توضیحات را فراهم می‌کند. این ویژگی، چارچوب پیشنهادی را به یک سازوکار قابل‌کنترل برای طراحی مدل‌هایی با کارایی بالا و توضیح‌پذیری تقویت‌شده تبدیل می‌نماید.

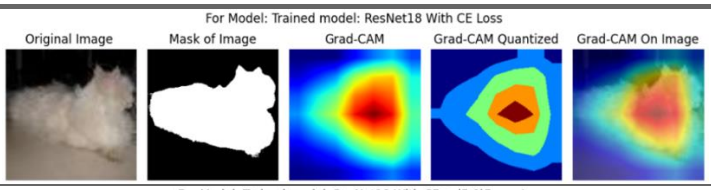
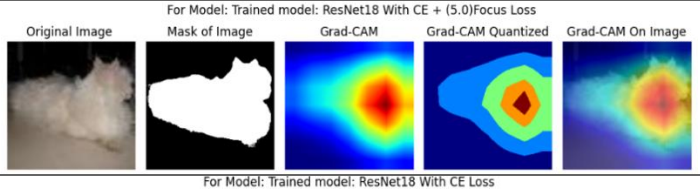
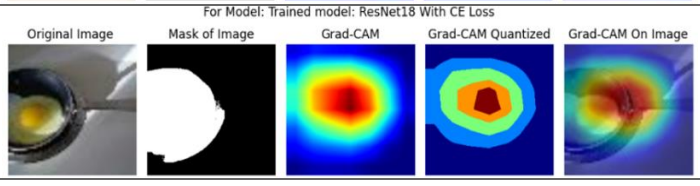
۳-۲. تحلیل نقشه حرارتی‌های مدل آموزش داده شده

پس از بررسی و ارزیابی نتایج طبقه‌بندی مدل پایه و مدل آموزش داده شده با روش پیشنهادی، برای بررسی کیفی نتایج توضیح‌پذیری نقشه حرارتی‌های تولید شده توسط هر مدل نیز بررسی و ارزیابی می‌شوند. در صورتی که نقشه حرارتی‌های تولید شده از روش پیشنهادی بهتر و معنادارتر از روش پایه باشند می‌توان با اطمینان بیشتری به نتایج مدل اعتماد پیدا کرد. در جدول ۳ نقشه حرارتی‌های تولید شده توسط مدل پایه و مدل‌های برتر آموزش داده شده با روش پیشنهادی مقایسه شده‌اند. برای نمایش تاثیر فرآیند بازآموزش، نقشه حرارتی تولید شده با مدل پایه نیز قرار داده شده است تا تاثیر کلی باز آموزش بر روی شبکه نیز نمایش داده شود و بتوان درک بهتری از تاثیر روش ارائه شده داشت.

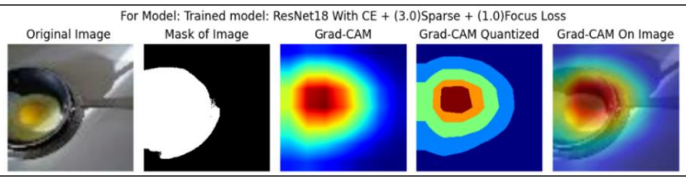
برای درک بهتر نقشه‌های حرارتی در جدول ۳، این نقشه در سه تصویر به نمایش در آمده است که توضیحات هر یک به این صورت است:

- تصویر با عنوان Grad-CAM، همان نقشه حرارتی تولید شده با روش Grad-CAM است.
- تصویر با عنوان Grad-CAM Quantized نقشه حرارتی Grad-CAM است با این تفاوت که مقادیر در چند بازه به صورت زیر کوانتیده شده‌اند:
  - مقدار پیکسل در بازه [۰,۰.۲۵] مقدار ۰ دارد
  - مقدار پیکسل در بازه [۰.۲۵,۰.۵] مقدار ۰.۲۵ دارد
  - مقدار پیکسل در بازه [۰.۵,۰.۷۵] مقدار ۰.۵ دارد
  - مقدار پیکسل در بازه [۰.۷۵,۰.۹] مقدار ۰.۷۵ دارد
  - مقدار پیکسل در بازه [۰.۹,۱.۰] مقدار ۱.۰ دارد
- تصویر با عنوان Grad-CAM On Image نقشه حرارتی تولید شده بر روی تصویر ورودی به مدل است.

جدول ۳.مقایسه نقشه حرارتی‌های تولید شده از مدل‌های برتر با مدل پایه و مدل از پیش آموزش داده شده.

تابع ضرر	β_{focus}	γ_{sparse}	Test Accuracy	نقشه حرارتی تولید شده پس از آموزش مدل
$\mathcal{L}_{CE}$	۰,۰	۰,۰	۸۹,۰٪	
$\mathcal{L}_{CE} + \gamma_{sparse}\mathcal{L}_{Sparse} + \beta_{focus}\mathcal{L}_{focus}$	۵,۰	۰,۰	۹۲,۸٪	
$\mathcal{L}_{CE}$	۰,۰	۰,۰	۸۹,۰٪	

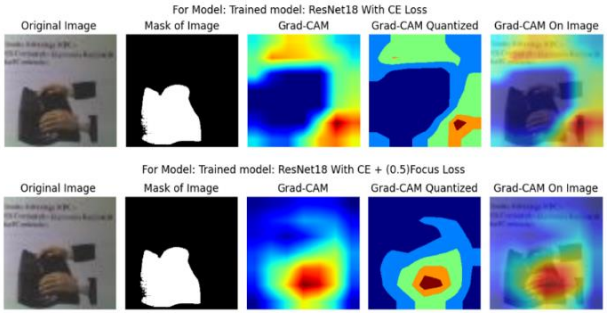
[Downloaded from ic4i-journal.ir on 2026-11-08]

$\mathcal{L}_{CE}$ $+ \gamma_{sparse}\mathcal{L}_{Sparse}$ $+ \beta_{focus}\mathcal{L}_{focus}$	۱,۰	۳,۰	۹۳,۸٪	

طبق نقشه‌های حرارتی جدول ۳، مدل‌های توضیح‌پذیر آموزش داده شده نسبت به مدل پایه نقشه‌های حرارتی با پوشش بیشتری از سوژه اصلی تولید کرده‌اند و هدف اصلی روش ارائه شده که افزایش تمرکز مدل بر روی سوژه اصلی و کاهش تاثیر سایر ویژگی‌های کم اهمیت ورودی را نشان می‌دهند. برای تحلیل نقشه‌های حرارتی می‌توان موارد زیر را به عنوان سنجه‌های ارزیابی مدنظر قرار داد:

- تمرکز نقشه حرارتی تولید شده بر سوژه اصلی و میزان پوشش سوژه اصلی با سطح‌های تاثیر بالاتر (مطلوب است بخش بیشتری از سوژه اصلی تصویر با نقاط دارای مقدار بیشتر در نقشه حرارتی تکمیل شوند)
- میزان پوشش پس زمینه تصویر با سطح‌های تاثیر پایین‌تر (به این صورت که مطلوب است بخش بیشتری از پس زمینه تصویر با نقاط دارای مقدار کمتر در نقشه حرارتی تکمیل شوند)

برای بررسی دقیق‌تر نقشه‌های حرارتی تولید شده دو نمونه دیگر از نقشه حرارتی‌های تولید شده توسط مدل پایه و روش پیشنهادی در شکل ٦ نشان داده شده است.



شکل ٦. تفاوت نقشه حرارتی تولید شده توسط مدل پایه و روش پیشنهادی.

طبق شکل ٦ و نتایج بررسی شده در جدول ۳، می‌توان تاثیر آموزش مدل با رویکرد پیشنهادی را به صورت زیر تحلیل کرد:

- روش پیشنهادی بر روی نمونه تصاویری که نقشه حرارتی نسبتا مناسبی با مدل پایه دارند، به این صورت است که، موجب تمرکز بیشتر بر روی سوژه اصلی تصویر شده و مدل را مجبور می‌کند تا ویژگی‌هایی را انتخاب کند که در محدوده مورد نظر از تصویر هستند.
- همچنین روش پیشنهاد شده بر روی تصاویری که نقشه حرارتی مناسبی با مدل پایه ندارند نیز عملکرد بهتری دارد، به این صورت که، به دلیل وجود ماسک در روش پیشنهادی موجب تمرکز بیشتر مدل آموزش داده شده به نقاط مهم تصویر می‌شود و باعث تولید نقشه حرارتی‌های بهتری از نظر تمرکز بیشتر بر سوژه اصلی و پراکندگی کمتر نقاط پر اهمیت می‌شود. تفاوت این نتایج در نقشه حرارتی‌های شکل ٦ نشان داده شده است.

با توجه به آزمایشات انجام شده، روش پیشنهادی عملکرد مطلوبی در افزایش توضیح‌پذیری مدل ResNet۱۸ داشته است. این عملکرد در جنبه‌های افزایش دقت پیش‌بینی مدل، بهبود نقشه‌های حرارتی و تمرکز بیشتر مدل بر روی سوژه اصلی مشهود است.

۳-۳. مقایسه با پژوهش‌های پیشین

به منظور ارزیابی دقیق‌تر عملکرد روش پیشنهادی و بررسی میزان اثربخشی آن در مقایسه با سایر پژوهش‌های مرتبط، آزمایش‌های تکمیلی بر روی چند مجموعه‌داده‌ی استاندارد مورد استفاده در تحقیقات پیشین انجام شد. پژوهش‌هایی که برای مقایسه انتخاب شدند، عمدتاً از مدل‌های متداول بینایی کامپیوتر مانند ResNet، VGG یا DenseNet همراه با روش‌های توضیح‌پذیری پس‌پردازشی نظیر Grad-CAM و LIME استفاده کرده بودند. در این پژوهش، همان مجموعه‌داده‌هایی که در آن مطالعات به کار رفته بودند، شامل VLCS مورد استفاده قرار گرفتند و مدل پیشنهادی با توابع زیان ($\mathcal{L}_{focus}$) و ($\mathcal{L}_{Sparse}$) بر روی آنها بازآموزی و ارزیابی شد.

[Downloaded from ic4i-journal.ir on 2026-11-08]

نتایج کمی به‌دست‌آمده از مقایسه با پژوهش‌های موجود در جدول ٤ آورده شده است. همان‌طور که از این جدول مشخص است، روش پیشنهادی در تعدادی از معیارهای ارزیابی از جمله **Comp** و **Suff** نسبت به روش‌های موجود عملکرد بهتری داشته است. به عنوان مثال، بر روی مجموعه‌داده‌ی **VLCS**، معیار **Comp** که در واقع بیانگر میزان تاثیر ویژگی‌های انتخاب شده بر پیش‌بینی مدل است، در مدل آموزش داده شده با روش پیشنهادی ٦٠.٠٤ گزارش شد که نسبت به بهترین روش پیشین با مقدار ٢٨.٥٣، بهبودی معادل ٣٢ واحد را نشان می‌دهد. با توجه به مفهوم معیار **Comp** این بهبود در معیار ارزیابی **Comp** بیانگر تاثیر افزایش تمرکز و استفاده از رویکرد اصلی روش پیشنهادی بوده است که در نتایج کمی نیز به طور واضحی قابل مشاهده و ارزیابی است. همچنین این بهبود در شاخص **Suff** نیز قابل مشاهده است.

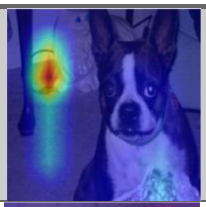
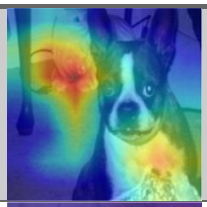
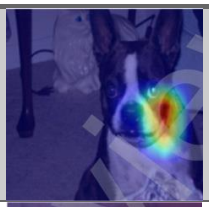
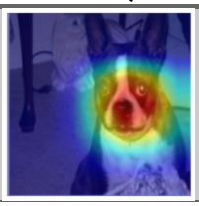
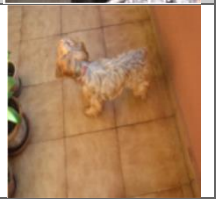
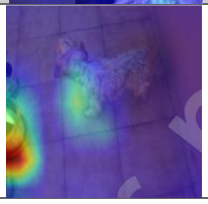
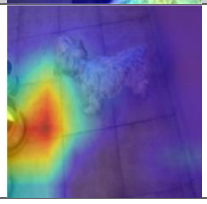
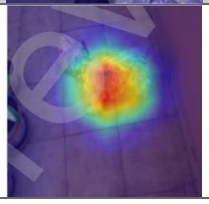
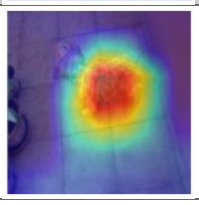
جدول ٤. مقایسه نتایج روش پیشنهادی با سایر پژوهش‌های انجام شده.

شبکه آموزش داده شده	Caltech١٠١			
	Accuracy	AR	Comp	Suff
ERM	٩٨,٦٨	٩٨,٣٤	٠,٤٨	٣٨,٦٩
IRM	٩٥,٣٥	٥٦,١٢	٢,٦٥	٣٠,٣١
GroupDRO	٩٧,٨٨	٥٠,٧٥	٥,٩١	٤٩,٦٨
Mixup	٩٩,٠٥	٥٨,٣٦	٧,٥٩	٣٧,٩٥
SGT	٩٩,٣٥	٩٨,٩٩	٠,٧٣	٤٧,٣٦
SGDrop	٩٩,٣١	٩٩,٤١	١٠,٠٩	٢٧,٣٣
DRE	٩٩,٢٤	٩٨,٥٥	٣,٢٥	٢٩,١٩
روش FocusEx	٩٩,٩٨	٩٩,٨٣	٢٨,٥٣	٢٨,٢٢
روش پیشنهادی	٩٧,١	٢٨,٧٤	٦٠,٠٤	٢,١٧

به‌ویژه در معیارهای کیفی مربوط به تمرکز مدل بر نواحی معنادار تصویر و کاهش پراکندگی نقشه‌های حرارتی.

علاوه بر نتایج کمّی، در معیارهای کیفی مربوط به تمرکز مدل بر نواحی معنادار تصویر و کاهش پراکندگی نقشه‌های حرارتی نیز روش پیشنهادی بهبود بسیار خوبی داشته است. تحلیل کیفی نقشه‌های حرارتی تولیدشده در جدول ٥ نشان می‌دهد که مدل پیشنهادی قادر است نواحی مرتبط‌تری از تصویر را شناسایی و بر بخش بیشتری از سوژه اصلی تمرکز کند. در مقایسه با روش‌های پیشین، نقشه‌های حرارتی حاصل از روش پیشنهادی تمرکز بالاتری و پوشش بیشتری از سوژه اصلی را داشتند. این موضوع بیانگر آن است که استفاده از روش پیشنهاد شده موجب هدایت مؤثرتر فرآیند یادگیری شبکه به سمت ویژگی‌های کلیدی و مؤثر تصویر شده است.

جدول ٥. مقایسه نقشه حرارتی تولید شده توسط روش پیشنهادی و سایر پژوهش‌های صورت گرفته بر روی تصاویر مجموعه داده VLCS.

تصویر اصلی	روش DRE	روش ERM	روش FocusEx	روش پیشنهادی
				
				

در مجموع، مقایسه با پژوهش‌های پیشین نشان می‌دهد که روش پیشنهادی علاوه بر بهبود معیارهای کمّی توضیح‌پذیری، از نظر جنبه‌های کیفی نیز خروجی‌های معنادارتری تولید کرده است. این نتایج بیانگر آن است که ترکیب رویکردهای یادگیری توضیح‌پذیر با طراحی تابع زیان هدف‌دار می‌تواند گامی مؤثر در جهت توسعه مدل‌هایی باشد که نه‌تنها دقیق‌تر، بلکه قابل اعتمادتر و قابل تفسیرتر هستند.

٤. بحث و نتیجه گیری

فصلنامهٔ علمی - پژوهشی فرماندهی و کنترل، سال دهم، شمارهٔ یک، بهار ١٤٠٥

[Downloaded from ic4i-journal.ir on 2026-11-08]

در این پژوهش، یک رویکرد نوین در زمینه یادگیری مبتنی بر توضیح‌پذیری (EGL)، برای بهبود توضیح‌پذیری شبکه‌های عصبی عمیق از طریق طراحی توابع زیان $\mathcal{L}_{focus}$ و $\mathcal{L}_{Sparse}$ پیشنهاد و ارزیابی شد. روش پیشنهادی بر روی شبکه‌ی ResNet۱۸ با وزن‌های از پیش آموزش‌دیده بر روی مجموعه‌داده ImageNet پیاده‌سازی شده است. اهدف اصلی این روش شامل موارد زیر است:

- هدایت مدل به سمت تمرکز بر نواحی معنادار تصویر و افزایش تمرکز مدل بر روی سوژه اصلی
- کاهش وابستگی مدل به ویژگی‌های غیرضروری یا پراکنده

آزمایش‌ها برای ارزیابی عملکرد در دو فاز انجام گرفتند تا تأثیر دقیق راهنمایی‌های توضیح‌پذیری بر کیفیت یادگیری مدل به‌صورت جامع بررسی شود. ارزیابی عملکرد طبقه‌بندی و تحلیل توضیح‌پذیری مدل.

نتایج به‌دست‌آمده نشان دادند که استفاده از ترکیب زیان‌های $\mathcal{L}_{focus}$ و $\mathcal{L}_{Sparse}$ در فاز بازآموزی مدل از پیش آموزش‌دیده، موجب بهبود قابل‌توجهی در عملکرد کلی شبکه می‌شود. این افزایش دقت بیانگر آن است که افزودن راهنمایی‌های توضیح‌پذیری با روش ارائه پیشنهادی نه تنها به درک بهتر ویژگی‌های مهم توسط مدل کمک کرده‌اند، بلکه منجر به بهبود قابلیت تعمیم و عملکرد کلی شبکه نیز شده‌اند.

در بخش تحلیل نقشه‌های حرارتی حاصل از مدل‌های مختلف، تفاوت معناداری میان مدل پایه و مدل‌های آموزش‌داده‌شده با روش پیشنهادی مشاهده شد. نقشه‌های حرارتی تولیدشده توسط مدل‌های توضیح‌پذیر، تمرکز بسیار بیشتری بر سوژه‌ی اصلی تصویر داشتند و نواحی پس‌زمینه با شدت پایین‌تری فعال شدند. این موضوع نشان می‌دهد که مدل توانسته است نواحی غیرمؤثر تصویر را بهتر نادیده گرفته و تصمیم‌گیری خود را بر اساس بخش‌های مرتبط‌تری از داده انجام دهد. علاوه بر این، در نمونه‌هایی که مدل پایه نتوانسته بود به‌درستی بر سوژه تمرکز کند، روش پیشنهادی توانست با هدایت مدل از طریق ماسک‌های تولیدی، تمرکز مکانی دقیق‌تر و منسجم‌تری ایجاد نماید.

همچنین، مقایسه با پژوهش‌های پیشین نشان می‌دهد که روش پیشنهادی نه‌تنها موجب بهبود معیارهای کمّی مرتبط با توضیح‌پذیری شده است، بلکه از نظر جنبه‌های کیفی نیز، نقشه‌های حرارتی منسجم‌تر و معنادارتری تولید کرده است. این نتایج تایید می‌کنند که ترکیب رویکردهای یادگیری توضیح‌پذیر با طراحی هدف‌مند توابع زیان می‌تواند به توسعه‌ی مدل‌هایی منجر شود که علاوه بر دقت بیشتر، قابلیت تفسیر و اعتمادپذیری بالاتری نیز دارند.

٤-١. نتیجه گیری

به‌طور کلی، یافته‌های این پژوهش تأیید می‌کنند که استفاده از توابع زیان مبتنی بر توضیح‌پذیری در قالب زیان‌های $\mathcal{L}_{focus}$ و $\mathcal{L}_{Sparse}$ می‌تواند به‌عنوان ابزاری مؤثر برای بهبود همزمان دقت و قابلیت تبیین در مدل‌های بینایی عمیق به‌کار گرفته شود. روش پیشنهادی عملکرد قابل قبولی در افزایش تمرکز نقشه‌های حرارتی تولیدشده داشته و نه‌تنها منجر به تولید نقشه‌های حرارتی معنادارتر شده است، بلکه رفتار تصمیم‌گیری منسجم‌تر و قابل‌اعتمادتر در مدل را نیز تقویت کرده است. در نتیجه، می‌توان گفت رویکرد پیشنهادی گامی مؤثر در جهت ترکیب آموزش هدف‌محور با تبیین‌پذیری مدل‌ها بوده و می‌تواند به‌عنوان مسیری عمومی برای طراحی مدل‌های توضیح‌پذیرتر در کاربردهای مبتنی بر داده‌های تصویری مورد استفاده قرار گیرد.

نتایج این پژوهش نشان می‌دهد که ادغام قیود توضیح‌پذیری در فرآیند آموزش، نقش مهمی در کنترل رفتار مدل ایفا می‌کند؛ به‌گونه‌ای که شبکه عصبی به‌صورت هدفمند به نواحی معنادار تصویر هدایت شده و از اتکا به سرنخ‌های نامرتبط یا نویزی پرهیز می‌نماید. همچنین، نقشه‌های حرارتی تولیدشده به‌عنوان یک لایه‌ی ارتباطی میان سامانه‌ی هوشمند و کاربر انسانی عمل کرده و امکان تحلیل، پایش و اعتمادسازی در تصمیمات مدل را فراهم می‌سازند. از سوی دیگر، استفاده از توابع زیان پیشنهادی باعث ارتقای سطح هوش سیستم از طریق هم‌راستاسازی فرآیند استنتاج با منطق قابل تبیین شده و این امر، استقرار چنین مدل‌هایی را در سامانه‌های مبتنی بر محاسبات و تصمیم‌یارهای عملیاتی حساس، واقع‌بینانه‌تر می‌سازد. بنابراین، رویکرد ارائه‌شده می‌تواند به‌عنوان گامی مؤثر در جهت توسعه‌ی

[Downloaded from ic4i-journal.ir on 2026-11-08]

سامانه‌های بینایی کامپیوتری مبتنی بر شبکه‌های عصبی عمیق با قابلیت تصمیم‌گیری شفاف در نظر گرفته شده و این شبکه‌های عصبی برای استفاده در سامانه‌های فرماندهی و کنترل نظامی قابل‌کنترل و قابل اعتماد تلقی شود.

در پایان، می‌توان نتیجه گرفت که رویکرد پیشنهادی گامی مؤثر در جهت ترکیب آموزش هدف‌محور با تبیین‌پذیری مدل‌ها بوده و می‌تواند به عنوان یک مسیر عمومی برای طراحی مدل‌های توضیح‌پذیرتر در کاربردهای مبتنی بر داده‌های تصویری مورد استفاده قرار گیرد. برای تبیین ادامه این مسیر تحقیقاتی موارد زیر از جمله بخش‌هایی هستند که موجب تاثیر بسزایی افزایش توضیح‌پذیری می‌شوند.

- آزمون و بررسی توابع ریاضیاتی دیگر برای هم‌راستایی ماسک و نقشه حرارتی
- تولید ماسک تصاویر با استفاده از روشی با اطمینان بالاتر و بدون نیاز به دستور

گسترش بحث یادگیری مبتنی بر توضیح‌پذیری برای سایر کاربردهای یادگیری‌عمیق شامل متن، صوت و سیگنال.

۵. مراجع

[۱] Y. Gao, S. Gu, J. Jiang, S. R. Hong, D. Yu, and L. Zhao (Apr. ۲۰۲٤), “Going Beyond XAI: A Systematic Survey for Explanation-Guided Learning”, ACM Comput Surv, Vol ٥٦, No ۷, PP ۱۸۸:۱-۱۸۸:۳۹, , doi: ۱۰,۱۱٤٥/۳٦٤٤۰۷۳.

[۲] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, and N. C. Bouaynaya (July ۲۰۲۲), “Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks,” IEEE Signal Process. Mag., Vol ۳۹, No ٤, PP ۷۳–۸٤, doi: ۱۰.۱۱۰۹/□□□.۲۰۲۲.۳۱٤۲۷۱۹.

[۳] A. Hedström et al. (Apr. ۲۸, ۲۰۲۳), “Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond”, arXiv: arXiv:۲۲۰۲,۰٦۸٦۱. doi: ۱۰,٤۸٥٥۰/arXiv.۲۲۰۲,۰٦۸٦۱.

[٤] Chen, Yanhui, Chao Chen, Xujia Li, Dongsheng Hong, and Xiangwen Liao. (Oct. ۲٤, ۲۰۲٥), "Focusex: Generalizable Explanation Guided Learning Through Noise Reduction and Fine-Grained Localization." *Available at SSRN 5403119*.

[٥] Shu, Shelley Zixin, et al. (۲۰۲٥), "Hybrid Explanation-Guided Learning for Transformer-Based Chest X-Ray Diagnosis." *International Workshop on Interpretability of Machine Intelligence in Medical Image Computing*. Cham: Springer Nature Switzerland.

[٦] Zeng, Guohang, et al. (۲۰۲۱), "Generating deep networks explanations with robust attribution alignment." *Asian Conference on Machine Learning*. PMLR.

[۷] Mitsuhara, Masahiro, et al. (۲۰۱۹), "Embedding human knowledge into deep neural network via attention map." *arXiv preprint arXiv:1905.03540*.

[۸] Ismail, Aya Abdelsalam, Hector Corrada Bravo, and Soheil Feizi. (۲۰۲۱), "Improving deep learning interpretability by saliency guided training." *Advances in Neural Information Processing Systems* ۳٤: ۲٦۷۲٦-۲٦۷۳۹.

[۹] DeYoung, Jay, et al. (۲۰۲۰), "ERASER: A benchmark to evaluate rationalized NLP models." *Proceedings of the 58th annual meeting of the association for computational linguistics*.

[۱۰] Chrysostomou, George, and Nikolaos Aletras. (۲۰۲۱), "Enjoy the salience: Towards better transformer-based faithful explanations with word salience." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

[۱۱] Atanasova, Pepa, et al. (۲۰۲۲), "Diagnostics-guided explanation generation." *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. ۳٦. No. ۱۰.

[۱۲] Situ, Xuelin, et al. (۲۰۲۱), "Learning to explain: Generating stable explanations fast." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

[۱۳] Ravi, Nikhila, et al. (۲۰۲٥), "Sam ۲: Segment anything in images and videos." *International Conference on Learning Representations*. Vol. ۲۰۲٥

[۱٤] Le, Yann, and Xuan Yang. (۲۰۱٥), "Tiny imagenet visual recognition challenge." *CS 231N* ۷,۷ : ۳.

[Downloaded from ic4i-journal.ir on 2026-11-08]